UNIVERSITY OF CALIFORNIA SAN DIEGO

Metabolic analysis of single cell gene expression data: What can we learn?

A Thesis submitted in partial satisfaction of the requirements for the degree
Master of Science

in

Bioengineering

by

Yuchen Zhou

Committee in Charge:

Bernhard Ø. Palsson, Chair
Prashant Gulab Ram Mali
Sheng Zhong
Daniel Zielinski

2020

The Thesis of Yuchen Zhou is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2020

DEDICATION

I would like to acknowledge Professor Bernhard Ø. Palsson for his support as the chair of my committee.

I would also like to acknowledge Dr. Daniel Zielinski for his mentoring. During one and a half year's research, his guidance and support has proved to be invaluable.

I would also like to acknowledge Professor Kun Zhang and the scientists working in his laboratory for providing us valuable datasets from experiment.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT OF THE THESIS

Metabolic analysis of single cell gene expression data:
What can we learn?

by

Yuchen Zhou

Master of Science in Bioengineering

University of California San Diego, 2020

Professor Bernhard Ø. Palsson, Chair

The advent of single cell profiling technologies has brought unprecedented resolution to cell heterogeneity in key human tissues. A significant remaining challenge is to interpret this cell variation in terms of meaningful functional differences and interactions between cell types. In this study, we perform a metabolic reconstruction-

based assessment of transcriptional heterogeneity in the human brain and kidney. We focus specifically on transporters as their expression is associated with both metabolic interactions between cell types and uptake of drugs. We find that: 1) we are able to identify drug-transporter relationships through structural homology between drugs and native transporter substrates, 2) we observe concomitant brain-region specific expression differences in key transporters that may impact therapeutic impact, 3) upstream metabolic genes have expression that correlates with the transporter expression, suggesting that transporter differences are likely physiologically significant, 4) examination of kidney data shows differential expression across different regions and cell types in the human kidney, 5) metabolic reconstructions provide useful interpretation to help understand the significance of single cell gene expression differences. Native metabolic activity of transporters is postulated through expression of metabolic genes with activities that are correlated as determined by metabolic flux modeling. This work illustrates the types of higher order interactions that can be elucidated from single cell profiling data and paves the way to clinical interventions targeting particular cell types and cell interactions.

INTRODUCTION

With rapid improvements to single-nucleus RNA sequencing methods, we are now gaining an increasingly clear picture of the differences in gene expression data between single cells in human tissues (Lake et al. 2017, 2016). Large scale single-cell expression data sets are now available for the human brain (Lake et al. 2016) and kidney (Lake et al. 2019). These studies have novel cell subtypes that exist in specific organ regions with defined physiological functions. The natural subsequent question is to determine whether differential expression of particular genes in certain cell subtypes can be linked to the phenotypes exhibited by these organ regions, which may include such features as healthy physiological functions or response to drug treatment.

There are many methods to understand high-dimensional biological data, for example statistical analysis and pathway analysis. One of the most successful methods for linking statistical differences to organism functions has been through the use of metabolic reconstructions. Metabolic reconstructions are carefully curated knowledgebase consisting of all known metabolic functions within an organism (Thiele and Palsson, 2010). These reconstructions serve as a platform to integrate various disparate data types to relate them quantitatively through network structure and function. The human metabolic network has been reconstructed and gradually improved over several iterations, culminating in the latest version Recon3D (Brunk et al. 2018).

Analysis of transporters specifically has a number of advantages. In the human brain and kidney, transporter activities, either neurotransmitter exchanges or filtration activities, are critical organ functions. Furthermore, it has been proposed that

transporters, as opposed to passive diffusion, may be responsible for a significant amount of drug uptake (O Hagan et al. 2015; Kell 2016). It was shown that similarities between endogenous metabolites in human tissues and drugs can be established via molecular fingerprints. Based on transporter expression, we may become able to determine region and cell specific targeting by drugs based on similarity to native compounds expressed by different cells. Last, we may begin to hypothesize metabolic interactions between different cell types co-localized in specific tissue or organs.

In this study, we collect native metabolite structures for all reconstructed human metabolites and single cell gene expression data for both human brain and kidney. Data on drug structures and known transporters are also collected in full from DrugBank (Wishart et al. 2008). We then compare pairwise structure similarity between natural metabolites and marketed drugs. Transporter expression data in each cell type are analyzed via machine learning methods for both the human brain and kidney. We then identify associations between particular transporters, metabolic functions, and likely region-specific drug uptakes. This work seeks to establish how metabolic reconstructions and chemical structural data can be combined with single-cell expression data to generate new hypotheses about the physiological role of newly identified cell subtypes.

METHODS

2.1. Extraction of transporter information from a human metabolic network reconstruction

To get the list of all transporter reactions, we use the Recon3D model from BiGG (bigg.ucsd.edu), which is the latest human metabolic network reconstruction. Then by using the Constraints-Based Reconstruction and Analysis (COBRA) approach, we extract a list of all transporter reactions that involve a same metabolite in both the extracellular compartment and the cytosol compartment from the Recon3D model. This transporter reactions list becomes the basis of the next steps. From the Recon3D model, the mapping of reactions, metabolites, and metabolite structures to the transporters is generated.

2.2. Extraction of drug information from DrugBank

On the other hand, the structures of all marketed drugs are obtained from DrugBank (drugbank.ca). We download the full drug structure links file that includes drug IDs, names, groups, proved transporters, and different systems of drug structures including general chemical formula, International Chemical Identifier (InChI), and Simplified Molecular-input Line-entry System (SMILES). To keep consistency and go through appropriate analysis, we use canonical SMILES structure from PubChem (pubchem.ncbi.nlm.nih.gov) for each metabolite and drug.

2.3. Calculation of structural similarity between drugs and native metabolites

Based on the structure links, we are able to calculate the similarity pairwise between native compounds and drugs based on Tanimoto coefficient of chemical fingerprints using RDKit, a universal toolkit for cheminformatics. A total of 1347 native compounds involved in transporter reactions and 9288 drugs are compared pairwise. By this calculation, we generate a similarity matrix and are able to identify drugs that are more similar to metabolites and to see if they tend to be transported by the same transporters. The similarity matrix also helps to explore drug-metabolite-transporter associations.

2.4. Analysis of single cell gene expression data of human brain and kidney

We obtained the gene expression data on a single cell basis for the human brain and kidney from Professor Kun Zhang's laboratory at UCSD. Nuclei substances were extracted from pre-operated cryosection from different regions in human brain. Single-nucleus assays were then run to measure the expression values in each cell and collect cell-type-specific expression data sets. We proceeded to analyze the data using machine learning methods. In order to compare the similarity of the drug to compound among several different dimensions, such as drugs and compounds share a transporter versus not sharing a transporter, transporters that proved to transport a specific drug versus not transport that drug, and transporters in different families involved membrane reactions, we use box and whisker plot to show the comparisons. Also, to provide validation, we use receiver operating characteristic (ROC) curves to predict drug-

transporter associations in binary class based on similarity values. ROC curves examine the diagnostic ability of a binary classifier system, and a higher score implies higher diagnostic ability of the system. To see the trend of single cell gene expression in human brain and kidney, we plot the expression heatmaps between transporter genes and cell types using Matplotlib in python. We also analyze the data via scikit-learn packages (sklearn) to group and cluster the dataset, and plot graphs using Matplotlib. After trying k-mean, Principal Component Analysis (PCA), and t-Distributed Stochastic Neighbor Embedding (TSNE), and calculating silhoutte coefficients for k-mean to visualize high-dimensional datasets, we choose TSNE clustering method to illustrate the results, since it shows the most distinctive clustering results. For each dataset, we test perplexity from 1 to 100, and choose the most appropriate value for each graph.

2.5. Analysis of upstream metabolic gene expression

To combine the human brain gene expression data with human metabolic network reconstruction, we use Escher (escher.github.io) to visualize biological pathways with correlated genes, proteins, enzymes, and reactions. Recon3D model provides a full human metabolic network to generate the biological pathway map. We use metabolic modeling as our tools to identify pathways linked to transported metabolites with a focus on neurotransmitters that are clearly defined metabolic functions of brain cells. We also use metabolic modeling method to explore the correlation between excitatory and inhibitory neurotransmitter gene expression.

RESULTS

3.1. Analysis of drug-transporter associations

To explore drug-transporter associations, we calculate drug-metabolite structural similarities via Tanimoto coefficient, and then link the drugs to particular transporters through the native activity of these transporters. We hypothesize that higher similarity scores imply co-transportation of native compound and drug. We use two native metabolites, 2-oxoglutarate (akg) and triiodothyronine (triodthy), as examples, and calculate the similarity scores for all drugs that share a transporter to the metabolites, respectively, as shown in Figure 1. According to the figure, drugs and metabolites transported by different transporters have similar scores.

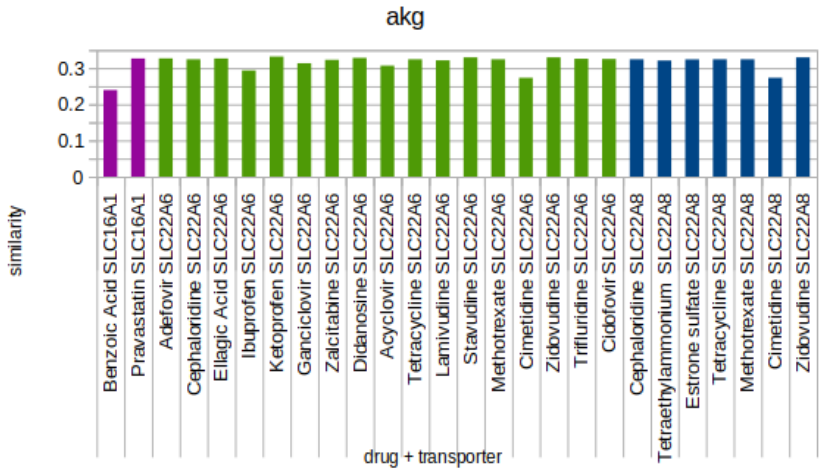

Figure 1.1: Tanimoto similarity scores of all the drugs that share transporters with metabolite akg (2-oxoglutarate), different transporters in different colors

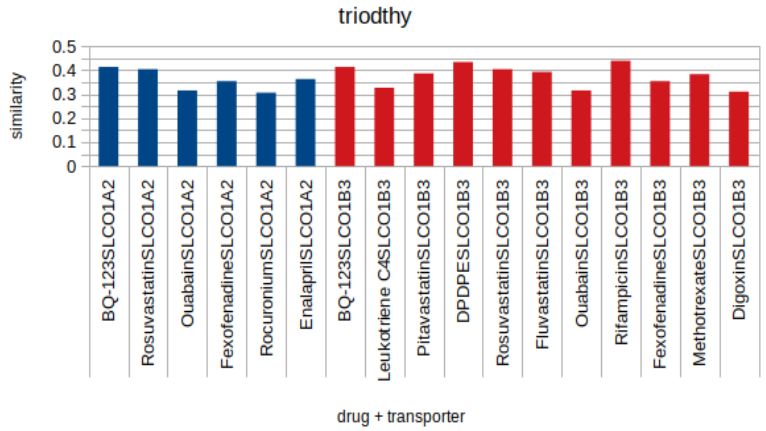

Figure 1.2: Tanimoto similarity scores of all the drugs that share transporters with metabolite triodthy (Triiodothyronine), different transporters in different colors

Figure 2 shows a set of box and whisker plots in order to broadly compare similarity of the drug to compounds that share a transporter versus not sharing a transporter. Figure 2.1 is the comparison of all the drugs sharing the transporter SLCO1B1 to all of the metabolites known to be transported by SLCO1B1, with each box indicating the average similarity value of all the metabolites compared to the specific drug. Figure 2.2 shows an example case of the drug Methotrexate, comparing all metabolites to the drug Methotrexate versus each transporter known to transport Methotrexate individually, each with the metabolites known to be transported by the transporter independently. With an average similarity value 0.37 of all the native compounds comparing to Methotrexate, the average similarity values of all the native compounds transported by transporter SLC22A11, SLC19A1, and SLC46A1, separately, are much higher, providing evidence that drugs and metabolites transported by these three transporters are more similar. The gene SLC19A1 encodes membrane

protein that serves as a transporter of folate and is involved in the regulation of intracellular concentrations of folate. This is consistent with the high similarity value 0.78 between methotrexate and folate. Figure 2.3 shows an example case of the drug Zidovudine, comparing all metabolites to Zidovudine versus each transporter known to transport Zidovudine individually, each with the metabolites known to be transported by the transporter independently. With an average similarity value 0.38 of all the native compounds comparing to Zidovudine, the average similarity values of all the native compounds transported by transporter SLC29A2 and SLC28A1, separately, are much higher, providing evidence that drugs and metabolites transported by these two transporters are more similar. The transporter SLC28A1 is associated with Acquired Immunodeficiency Syndrome (AIDS), and the drug Zidovudine is one of the treatments for AIDS. This result validates that highly similar metabolites and drugs can be co-transported by a transporter that has similar function to the drug.

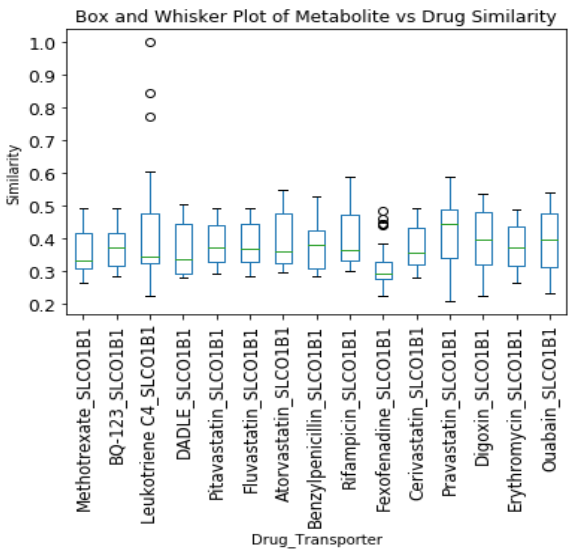

Figure 2.1: Comparison of the drugs sharing the transporter SLCO1B1 to all of the metabolites known to be transported by SLCO1B1

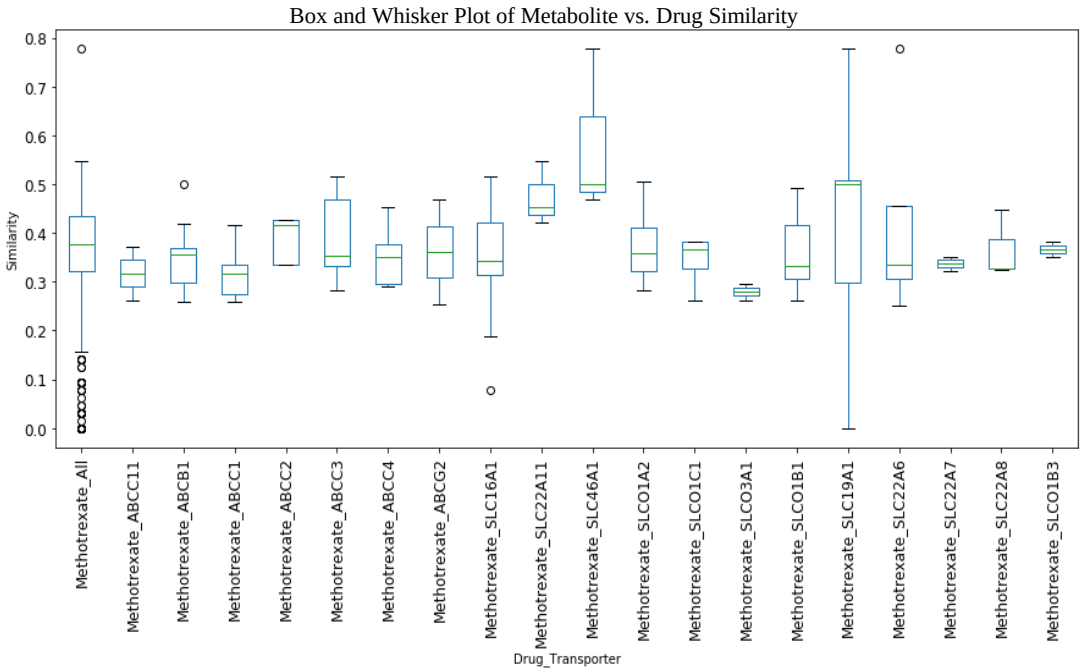

Figure 2.2: Comparison of all metabolites to Methotrexate versus each transporter known to transport Methotrexate individually, each with the metabolites known to be transported by the transporter independently

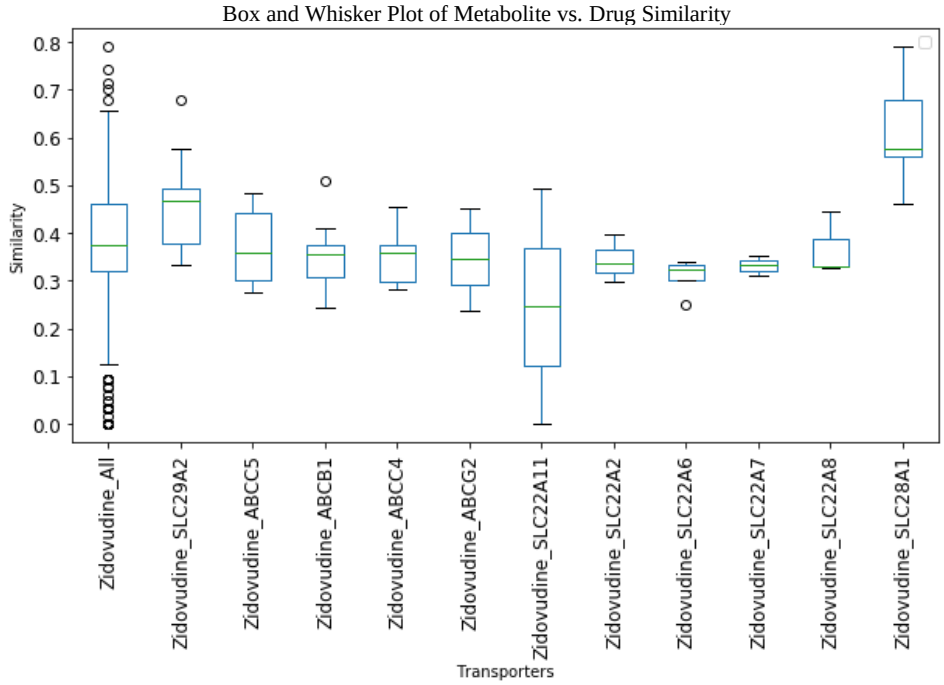

Figure 2.3: Comparison of all metabolites to Zidovudine versus each transporter known to transport Zidovudine individually, each with the metabolites known to be transported by the transporter independently

Then we compare transporters in solute carrier family (SLC) to transporters not in SLC family, each with the metabolites known to be transported by the transporter independently. Figure 2.4 indicates that the drug-metabolite similarity values are higher when transported by transporters in SLC family than in non-SLC family. An explanation is that SLC genes encode more membrane-bound proteins that facilitate more membrane transportations and have better co-transporting ability.

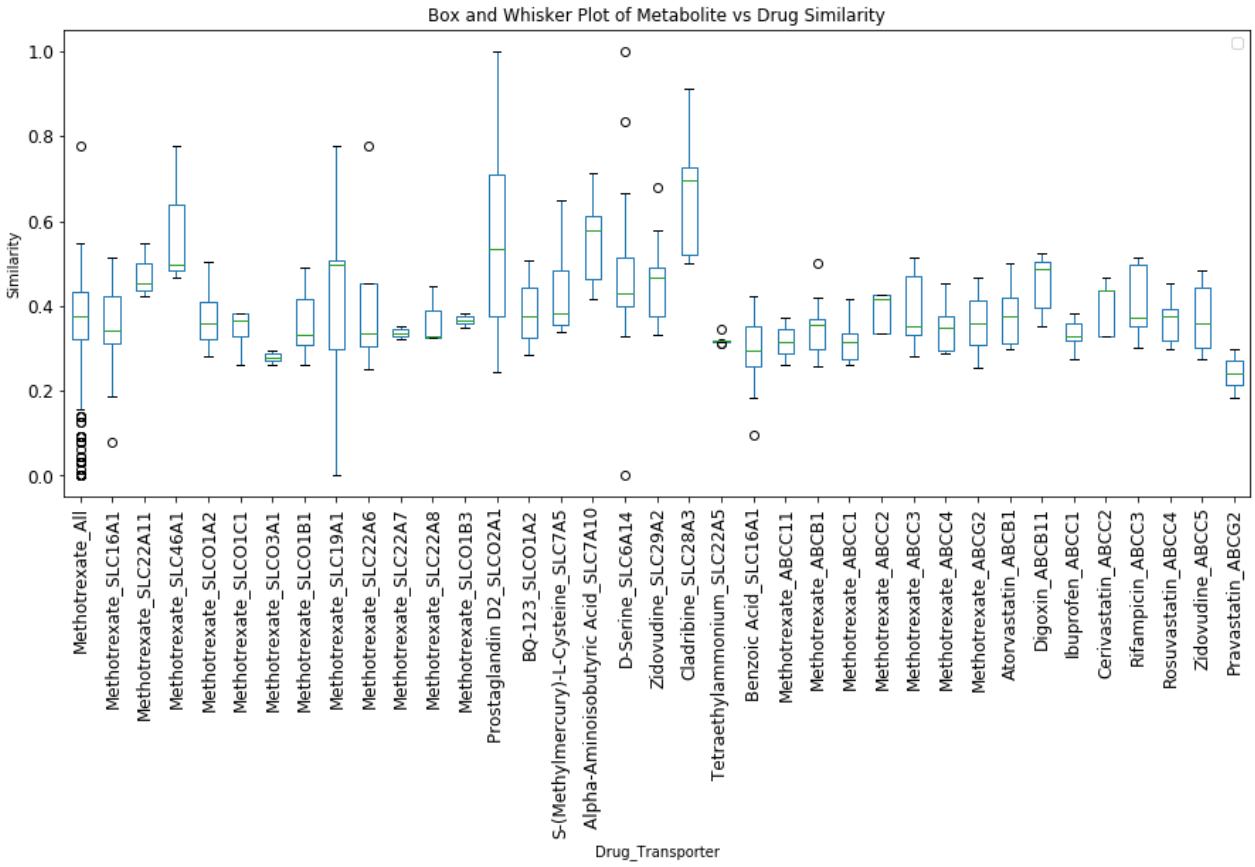

Figure 2.4: Comparison of transporters in SLC family to transporters not in SLC family, each with the metabolites known to be transported by the transporter independently

3.2. Prediction of drug-transporter relationships based on structural similarity

We use receiver operating characteristic (ROC) curves as a validation to predict drug-transporter associations based on similarity values and choose a cutoff value from 0.58 to 1. As shown in Figure 3, the area under the ROC curve is 0.85, proving that we are predicting the most likely transporters by comparing drugs to native metabolites.

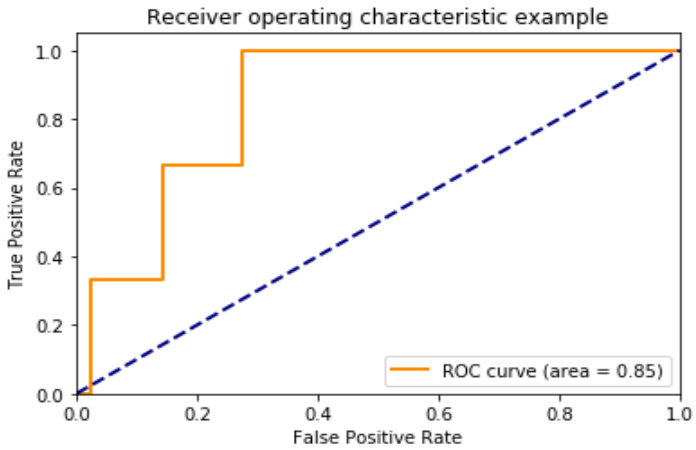

Figure 3.1: Receiver Operating Characteristic (ROC) curve for drug-metabolite associations based on Tanimoto Similarity values (drug-transporter pairs from supplementary table 1 of Dobson and Kell's paper, 2008)

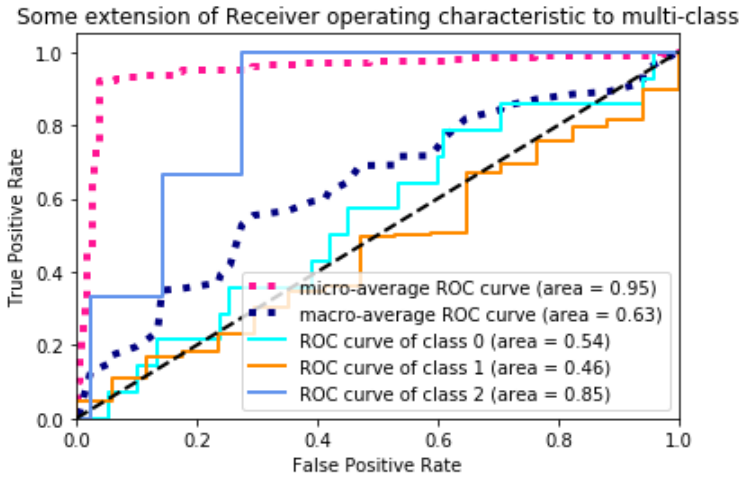

Figure 3.2: More ROC curves extended to multiclass

We also test drug transporter expression across cell types, and have four example drugs as case studies, including Methotrexate, Zidovudine, Benzylpenicillin, and Rosuvastatin. As shown in figure 4, for Methotrexate, cell End have the highest expression across transporter types, and transporter SLCO3A1 shows the highest expression across cell types, which has brain indication from DrugBank to treat Soft Tissue Sarcoma (STS). For Zidovudine, cell End also shows the highest expression value, and transporter ABCC5 has the highest expression value across cell types. For Benzylpenicillin, cell End and Mic have higher transporter expression, and transporter SLCO3A1 shows the highest expression across cell types. Lastly, for Rosuvastatin, cell End, Ast, and Oli has higher expression values, and transporter ABCC1 has the highest expression across cell types. This set of results indicates that compared to excitatory and inhibitory neurons in human brain, non-neural cells show much higher expression values of the transporter genes. Some transporter such as SLCO3A1, with high expression values in different types of brain cells, is proved to be efficient to treat specific brain-related diseases.

Figure 4.1: Average expression values of transporter genes known to transport Methotrexate across 21 cell types

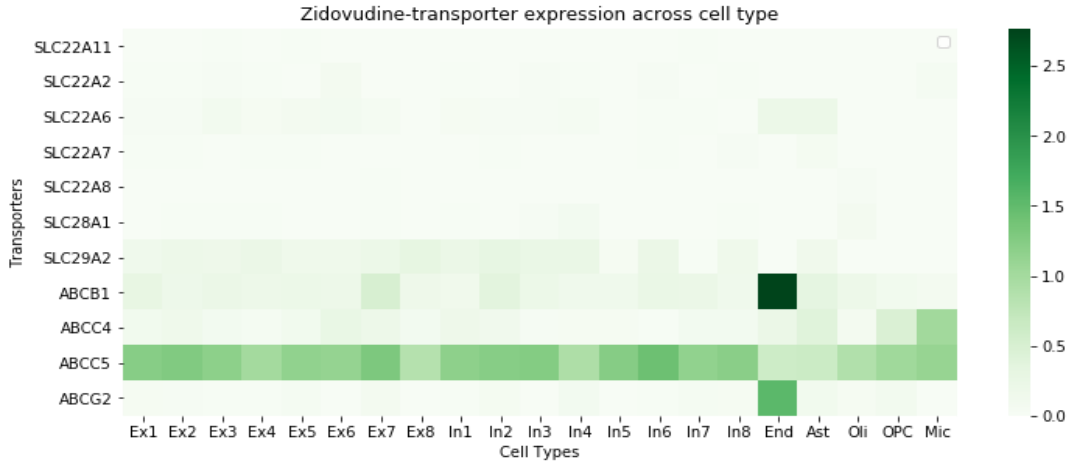

Figure 4.2: Average expression values of transporter genes known to transport Zidovudine across 21 cell types

Figure 4.3: Average expression values of transporter genes known to transport Benzylpenicillin across 21 cell types

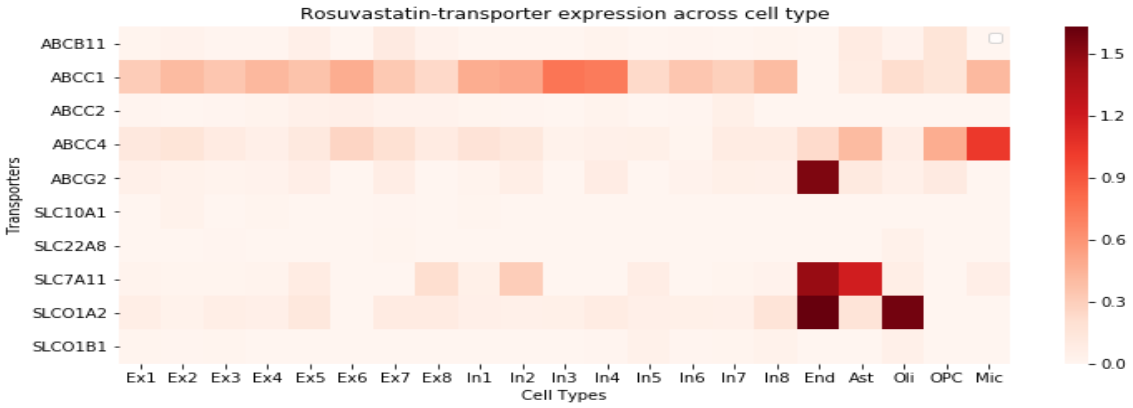

Figure 4.4: Average expression values of transporter genes known to transport Rosuvastatin across 21 cell types

3.3. Analysis of transporter gene expression in the human brain at single-cell resolution

We first calculate average expression values of transporter genes for 21 cell types. Figure 5 shows the simple mapping of expression data to transporters for each cell type. The resulting heatmaps indicate that cell type Ex5, Ast, and End have differentially expressed transporter data. We also use TSNE clustering method to illustrate transporter expression across different genes and cell types (Figure 6). Although the 21 cell types are functionally classified into three categories (excitatory neurons, inhibitory neurons, and non-neural cells), the clustering patterns based on transporter expression alone does not show a clear distinction between the cell types.

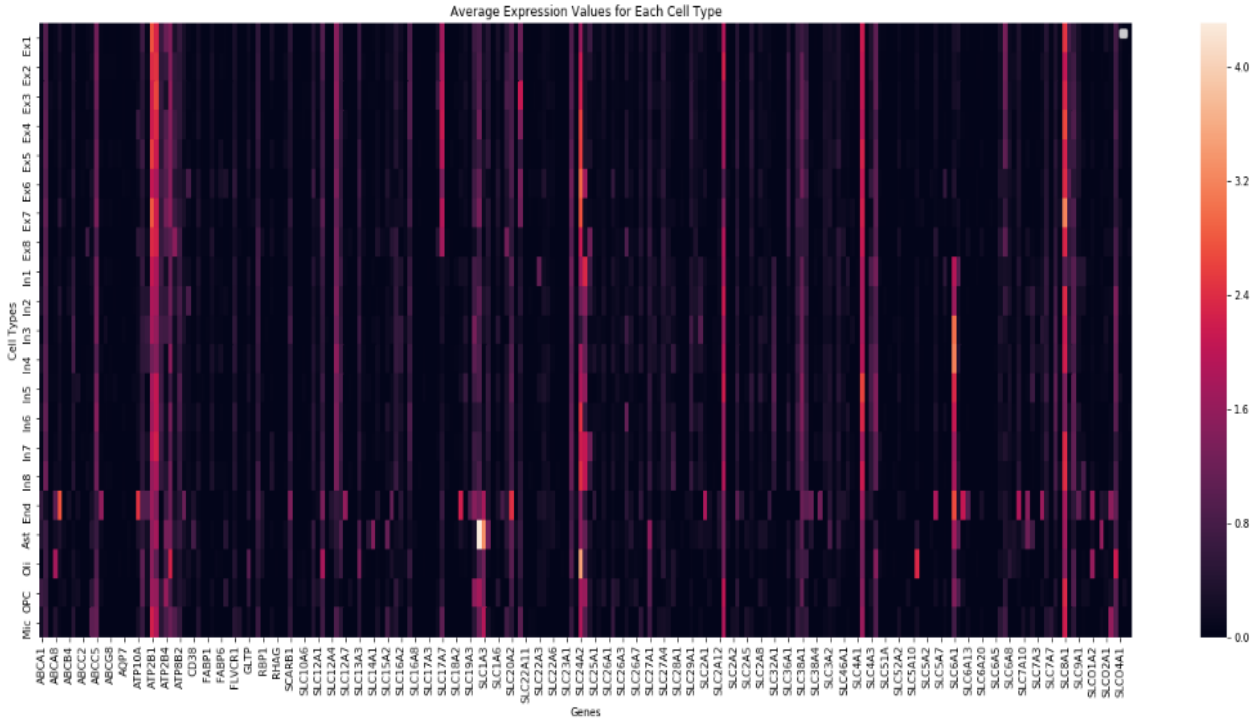

Figure 5: Average expression values of transporter genes for each brain cell type

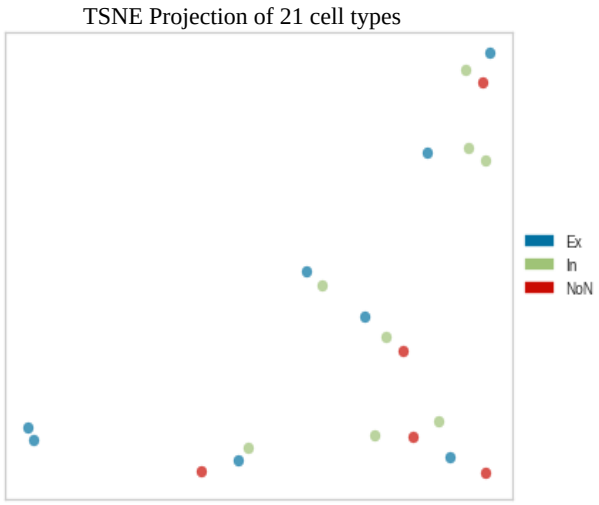

Figure 6.1: TSNE plot of 21 cell types (perplexity = 2)

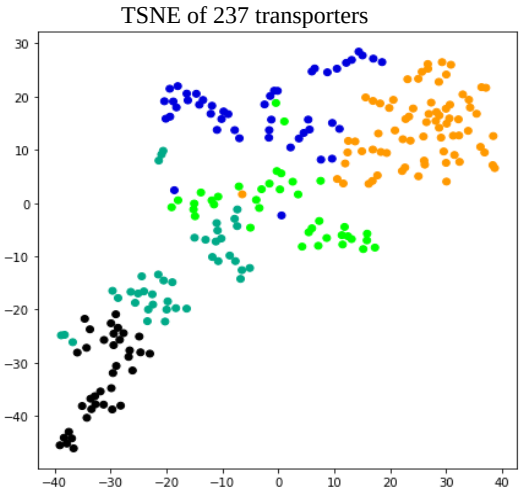

Figure 6.2: TSNE plot of 237 genes/transporters (perplexity = 10)

3.4. Analysis of single-cell gene expression data for the human kidney

We have similar analysis for kidney transporter expression and cell-specific metabolite transport data. The 30 cell types measured for expression data are from different functional structures in adult human kidney (Lake et al. 2019). Figure 7 collects the mapping and clustering results for the kidney data. After biclustering the data, a clear pattern shows up in Figure 7.2. We can easily classify the transporters into three

categories: variable expression, cell subtype-linked expression, and uniform expression pattern. By visiting the original data and kidney anatomy image, we can conclude that the transporters have higher expression values in proximal tubules, glomerular capsules, descending thin limb, and thin ascending limb.

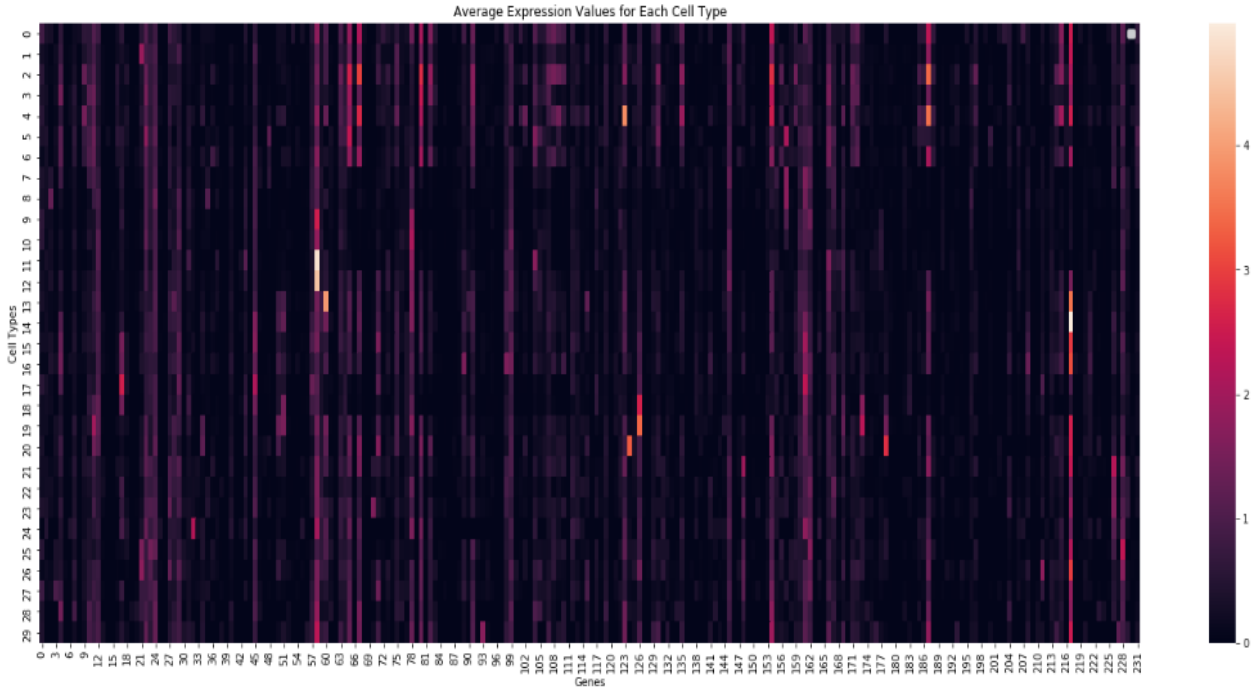

Figure 7.1: Average expression values of transporter genes for each kidney cell type (log scale)

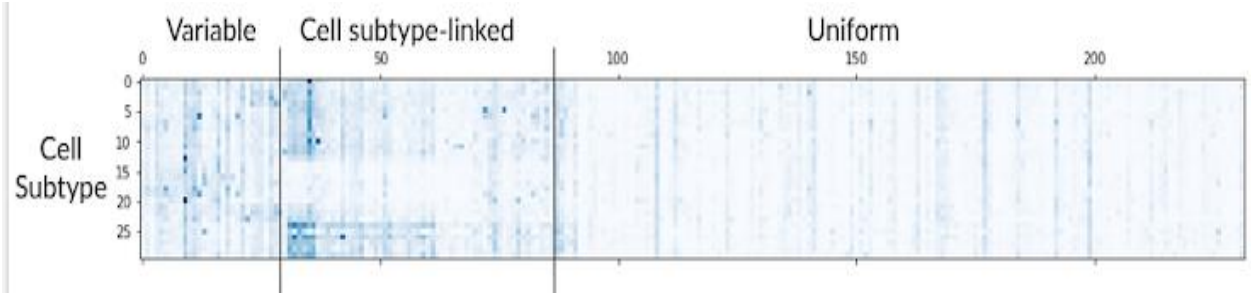

Figure 7.2: Biclustering structure of rearranged data

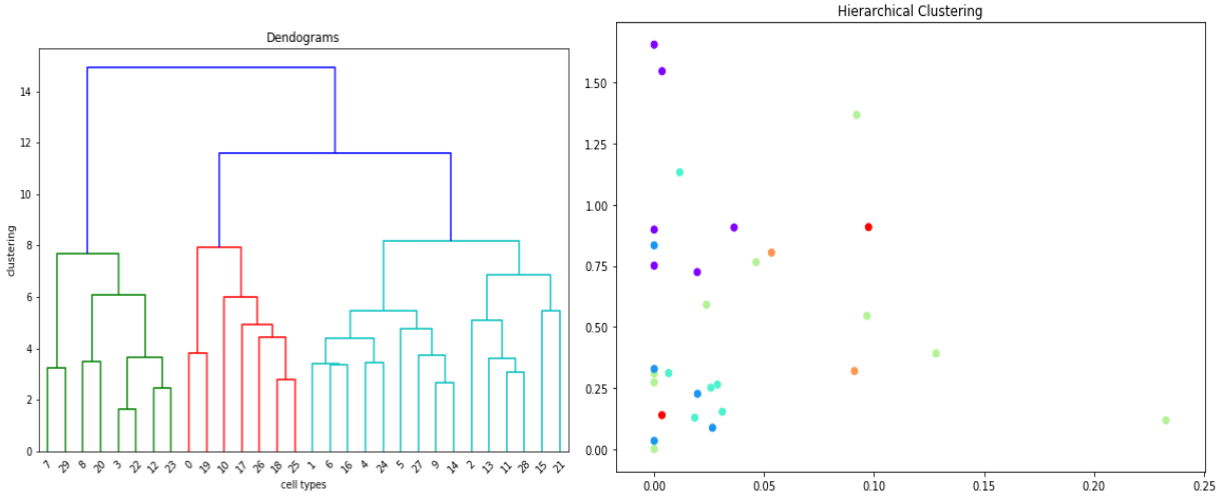

Figure 7.3: Hierarchical Clustering of kidney data (30 cell types, 6 clusters)

Due to the more complicated structure than human brain and various cell types in human kidney, the results are more complex for analyzing and identifying related kidney diseases. Despite the latest expression data for human kidney, scientists can hardly ethically gain normal kidney tissue source from human. With better tested data in the future, we can go further towards the molecular metabolic analysis and physiological organization of human kidney.

DISCUSSION

In this work, we sought to determine whether we could use metabolic network reconstructions and chemical structure information to empower the analysis of human single-cell gene expression data.

A key hypothesis underlying the analysis in this study was the postulation that transporters, rather than passive diffusion, may be responsible for the majority of drug uptake in tissues. We used similarity of drugs to native transported substrates to predict possible drug transporters. During this analysis, it became clear that assigning cutoffs to similarity was difficult. In comparison to experimentally validated drug transporters, certain transporters were validated while others did not mean a similarity cutoff to their native substrate. A key limiting factor was the small number of validated transporters, and importantly negative samples where a drug was established as not being transported. Such cases may help to establish relevant cutoffs for similarity required for shared transport activity.

Due to limited data on experimentally-determined drug transporters, assessment of transporters linked to drugs was dependent upon a structural comparison to native transporter substrates. It is clear that the chemical fingerprints, as well as the Tanimoto coefficient calculated thereafter, are not ideal metrics for predicting likelihood of sharing a transporter. The most successful predictions were the trivial cases where the drug is a substrate analog of the native transporter substrate. However, chemicals can share a transporter with much more diverse structures, for example, transporters that can recognize a broad array of amino acids with substantial structural differences. It is

possible that a moiety-based similarity metric may be more useful, but such a metric is more difficult to develop and requires its own testing (O'Hagan and Kell 2017).

Through integration of single-cell gene expression data with drug transporter predictions, we sought to identify subtypes of cells or regions of tissue that would be preferentially affected by certain drugs. This requires identification of transporter expression patterns tied to cell subtypes. Surprisingly, the cell subtypes did not clearly cluster into particular transporter expression patterns, despite largely sharing from a few neurotransmitter use patterns. It is possible that non-neurotransmitter-related transporters dominate the clustering, and expression of these genes is more heterogeneous. Further analysis of particular subtypes or regions affected by drugs was largely prospective; very little experimental data was found on regions of the brain affected by drugs. This may be a fruitful avenue for more investigation for example by comparing to fMRI analysis during drug treatment. Additionally, we wanted to identify native metabolic functions of these cell types. A complication here is the fact that many transporters can transport multiple substrates, thus the relevant metabolic activity is not clear. We can additionally examine expression of genes upstream from each possible transported native substrate, which has proven helpful when examining neurotransmitter pathways for example (Wandschneider and Koepp 2016). Finally, examination of kidney data showed differential expression across different regions and cell types. This contributes to further analysis of kidney anatomy and functional region identifications.

In summary, the work presented gives sufficient support for further examination of the integration of chemical structures, metabolic networks, and single-cell gene expression. A key next step will be gathering experimental data to validate predicted phenotypes tied to particular organ regions and cell subtypes, as well as histological relationships between cell subtypes that may better inform likely metabolic interactions.

REFERENCES

Brunk Elizabeth, Swagatika Sahoo, Daniel C. Zielinski, Ali Altunkaya, Andreas Dräger, Nathan Mih, Francesco Gatto, Avlant Nilsson, German Andres Preciat Gonzalez, Maike Kathrin Aurich, Andreas Prlić, Anand Sastry, Anna D. Danielsdottir, Almut Heinken, Alberto Noronha, Peter W. Rose, Stephen K. Burley, Ronan M.T. Fleming, Jens Nielsen, Ines Thiele, and Bernhard O. Palsson. 2018. “Recon3D Enables a Three-Dimensional View of Gene Variation in Human Metabolism.” *Nature Biotechnology* 36 (3): 272–81.

Kell Douglas B. 2016. “Implications of Endogenous Roles of Transporters for Drug Discovery: Hitchhiking and Metabolite-Likeness.” *Nature Reviews. Drug Discovery* 15 (2): 143.

Lake Blue B., Rizi Ai, Gwendolyn E. Kaeser, Neeraj S. Salathia, Yun C. Yung, Rui Liu, Andre Wildberg,Derek Gao, Ho-Lim Fung, Song Chen, Raakhee Vijayaraghavan, Julian Wong, Allison Chen, Xiaoyan Sheng, Fiona Kaper, Richard Shen, Mostafa Ronaghi, Jian-Bing Fan, Wei Wang, Jerold Chun, and Kun Zhang. 2016. “Neuronal Subtypes and Diversity Revealed by Single-Nucleus RNA Sequencing of the Human Brain.” *Science* 352 (6293): 1586–90.

Lake Blue B., Song Chen, Masato Hoshi, Nongluk Plongthongkum, Diane Salamon, Amanda Knoten, Anitha Vijayan, Ramakrishna Venkatesh, Eric H. Kim, Derek Gao, Joseph Gaut, Kun Zhang, and Sanjay Jain. 2019. “A Single-Nucleus RNA-Sequencing Pipeline to Decipher the Molecular Anatomy and Pathophysiology of Human Kidneys.” *Nature Communications* 10 (1): 2832.

Lake Blue B., Simone Codeluppi, Yun C. Yung, Derek Gao, Jerold Chun, Peter V. Kharchenko, Sten Linnarsson, and Kun Zhang. 2017. “A Comparative Strategy for Single-Nucleus and Single-Cell Transcriptomes Confirms Accuracy in Predicted Cell-Type Expression from Nuclear RNA.” *Scientific Reports* 7 (1): 1–8.

Paul D. Dobson, and Douglas B. Kell. 2008. “Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule?.” *Nat Rev Drug Discov* **7,** 205–220.

O’Hagan Steve, and Douglas B. Kell. 2017. “Analysis of Drug-Endogenous Human Metabolite Similarities in Terms of Their Maximum Common Substructures.” *Journal of Cheminformatics* 9 (March): 18.

O Hagan Steve, Neil Swainston, Julia Handl, and Douglas B. Kell. 2015. “A ‘Rule of 0.5’ for the Metabolite-Likeness of Approved Pharmaceutical Drugs.” *Metabolomics: Official Journal of the Metabolomic Society* 11 (2): 323–39.

Thiele Ines, and Bernhard Ø. Palsson. 2010. “A Protocol for Generating a High-Quality Genome-Scale Metabolic Reconstruction.” *Nature Protocols* 5 (1): 93–121.

Wandschneider Britta, and Matthias J. Koepp. 2016. “Pharmaco fMRI: Determining the Functional Anatomy of the Effects of Medication.” *NeuroImage. Clinical* 12 (October): 691–97.

Wishart David S., Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. 2008. “DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets.” *Nucleic Acids Research* 36 (Database issue): D901–6.